REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)* 01-04-2015	**2. REPORT TYPE** Final	**3. DATES COVERED** *(From - To)* 28-03-2013 – 27-03-2015

4. TITLE AND SUBTITLE Methods of sparse modeling and dimensionality reduction to deal with big data	**5a. CONTRACT NUMBER** FA2386-13-1-4046 **5b. GRANT NUMBER** Grant AOARD-134046 **5c. PROGRAM ELEMENT NUMBER** 61102F
6. AUTHOR(S) Tu Bao Ho	**5d. PROJECT NUMBER** **5e. TASK NUMBER** **5f. WORK UNIT NUMBER**
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Japan Advanced Institute of Science and Technology 1-1 Asahidai, Nomi, Ishikawa, 923-1292 1-2 Japan	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) AOARD UNIT 45002 APO AP 96338-5002	**10. SPONSOR/MONITOR'S ACRONYM(S)** AFRL/AFOSR/IOA(AOARD) **11. SPONSOR/MONITOR'S REPORT NUMBER(S)** AOARD-134046

12. DISTRIBUTION/AVAILABILITY STATEMENT

Distribution A: Approved for public release. Distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This project focused on the development of new methods for sparse modeling and non-negative matrix factorization. The major achievements are 1) a sparse topic model that can learn thousands of topics from a large set of documents and infer the topic mixture of each document, 2) a supervised dimension reduction method for large datasets, and 3) a non-negative matrix factorization (NMF) method with good interpretability. The research on sparse topic model and supervised dimension reduction as well as NMF is motivated by the need of reducing complexity in dealing with huge and complex datasets in big data. The proposed methods were theoretically and experimentally evaluated, and applied to problems in materials science and biomedicine.

15. SUBJECT TERMS

Sparse modeling, Dimensionality reduction, Non-negative matrix factorization, Probabilistic graphical model, Sparse topic model, Feature selection, Feature transformation, Dirichlet prior

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Hiroshi Motoda, Ph. D.
U	U	U	SAR	36	**19b. TELEPHONE NUMBER** *(Include area code)* +81-42-511-2011

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE
01 APR 2015

2. REPORT TYPE
Final

3. DATES COVERED
28-03-2013 to 27-03-2015

4. TITLE AND SUBTITLE
Methods of sparse modeling and dimensionality reduction to deal with big data

5a. CONTRACT NUMBER
FA2386-13-1-4046

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Tu Bao Ho

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Japan Advanced Institute of Science and Technology,1-1 Asahidai,Nomi, Ishikawa,923-1292 Japan,JP,9231292

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AOARD, UNIT 45002, APO, AP, 96338-5002

10. SPONSOR/MONITOR’S ACRONYM(S)
AFRL/AFOSR/IOA(AOARD)

11. SPONSOR/MONITOR’S REPORT NUMBER(S)
AOARD-

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This project focused on the development of new methods for sparse modeling and non-negative matrix factorization. The major achievements are 1) a sparse topic model that can learn thousands of topics from a large set of documents and infer the topic mixture of each document, 2) a supervised dimension reduction method for large datasets, and 3) a non-negative matrix factorization (NMF) method with good interpretability. The research on sparse topic model and supervised dimension reduction as well as NMF is motivated by the need of reducing complexity in dealing with huge and complex datasets in big data. The proposed methods were theoretically and experimentally evaluated, and applied to problems in materials science and biomedicine.

15. SUBJECT TERMS
Sparse modeling, Dimensionality reduction, Non-negative matrix factorization, Probabilistic graphical model, Sparse topic model, Feature selection, Feature transformation, Dirichlet prior

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE	**Same as Report (SAR)**	**36**	
unclassified	**unclassified**	**unclassified**			

Final Report for Grant AOARD-13-4046

Methods of sparse modeling and dimensionality reduction to deal with big data

Name of Principal Investigator: Tu Bao Ho

Period of Performance: March 28, 2013- March 27, 2015

Final Report for Grant AOARD-13-4046

Methods of sparse modeling and dimensionality reduction to deal with big data

Name of Principal Investigator: Tu Bao Ho

E-mail: bao@jaist.ac.jp
Institution: Japan Advanced Institute of Science and Technology
Mailing Address: JAIST, 1-1 Asahidai, Nomi City, Ishikawa, 923-1292 Japan
Phone and Fax: 81-761-51-1730

Period of Performance: 3/28/2013- 3/27/2015

Abtract

This project focuses on development of new methods for sparse modeling and non-negative matrix factorization, as their applicability.

In the first year, we focused on establishment of the methods: (i) a sparse topic model that can learn thousands of topics from a large set of documents and infer the topic mixture of each document, (i) a method of supervised dimension reduction for large datasets, and (iii) a non-negative matrix factorization (NMF) method with interpret-ability. The research on sparse topic model and supervised dimension reduction as well as NMF is motivated by the need of reducing complexity in dealing with huge and complex datasets in big data. The proposed methods were theoretically and experimentally evaluated.

In the second year, we continued to improve and test the methods as well as to employ sparse modeling and dimensionality reduction, including the developed methods, in our research in materials science and biomedicine.

The results of the project are reported as papers published in peer-reviewed journals and conferences.

Contents

1. Introduction

Big data are datasets that are very large and complex that current IT techniques are not able to deal well with them. The features of big data typically include: (a) Huge number of instances and variables (dimensions), (b) The complex and changing nature of the data. *Sparse modeling* and *dimensionality reduction* are two key approaches to reduce the data size and to manage complex relationships among data, thus they have greatly attracted the attention of machine learning research community.

Sparse modeling methods– aiming at using a small number of variables to model the problem– are largely based on Lasso and recently emerged as promising direction to big data. Typical methods include relaxed Lasso [15], bootstrap Lasso [16], group Lasso [17], etc. or beyond Lasso as sparse PCA [29], sparse NMF, etc.

Topic models– a case of probabilistic graphical models, which has been recently matured as marriage of probability theory and graph theory providing a powerful tool for modeling and solving problems related to uncertainty and complexity [20]. Typical topic modeling methods include probabilistic latent semantic analyzing (PLSA) [21] and latent Dirichlet allocation (LDA) [14], and in the last decade a large number of work has made a significant progress in topic model research. Most topic models developed so far are dense models that required all the words of the dictionary appear in each topic and all learned topics appeared in describing the original documents, and thus they cannot work for large datasets. Recently, there have been work on sparse topic models, typically regularized LSI [25], spase topical coding [30], etc. Even though these models provide elegant solutions to the sparsity problem, there remain some serious drawbacks when dealing with large-scale settings.

Nonnegative matrix factorization (NMF) is the problem of analyzing a original data matrix as product of two matrices (one presents the basis of the new feature space and the other presents the data represented in the new feature space). A recent comprehensive review of Wang [26] shows increasing fundamental role of NMF in transformations for dimension reduction and component analysis. For big data, approximate factorization, computing speed, sparsity solution and interpretability for NMF are being received more attention.

Sparse modeling and NMF are powerful approaches to *dimensionality reduction*, the key issue in big data analytics, and *applications* of big data analysis always require sparse modeling and dimensionality reduction. Concerning to all of the above mentioned, our project aims to do the followings:

1. **Task 1**: Develop a sparse topic model that can learn a large number of topics from a large collection of objects and represents each object by a small number of topics.
2. **Task 2**: Develop a supervised dimension reduction method and based on that a classification method that effectively works for tough situations such as non-linear separable cases or short sequences.
3. **Task 3**: Develop a new formulation of non-negative matrix factorization problem with high computation performance and a clear interpretation.
4. **Task 4**: Use sparse models and dimensionality reduction techniques in application, and develop related methods as well.

Concerning Task 1, we developed FSTM, a fully sparse topic model [6] that allows us to represent each topic by a subset of words in the dictionary and to represent each document as a mixture of a small number of learned learned. In order to deal with large scale topic models, we developed DOLDA [5], an online version of the Frank-Wolfe algorithm that allows us to first make an online version of LDA and following it an online version of FSTM has been developing.

Concerning Task 2, we developed SDR, a two-phase framework for doing dimension reduction of supervised discrete data [1]. The framework was demonstrated to exploit well label information and local structure of the training data to find a discriminative low-dimensional space.

Concerning Task 3, we developed sNMF (simplicial non-negative matrix factorization) [4], as a new formulation of NMF with constraint as a convex combination of latent components with significant properties such as interpretability, sparsity, high performance in classification task. To provide a basic tool for further development of NMF, we developed a fast and robust anti-lopsided algorithm for non-negative least squares (NNLS) with high accuracy [7].

Concerning Task 4, we developed a technique to analyze short sequences and applied it in biomedicine research [8] based on SDR and FSTM, and employed sparse modeling as well as dimensionality reduction in study of materials design [3].

In section 3, we will briefly present each of these papers obtained by the project.

2. The developed methods, their evaluation and discussion

In the followings, we will present each developed method, the theoretical and experimental evaluation, and discussion on its merit as well.

2.1 Sparse topic models

2.1.1 Fully sparse topic model (FSTM) [6]

The objective of this work is to develop a sparse topic model that learns *sparse topics* (i.e., topics represented by a proper subset of terms in of the vocabulary instead of all terms as in a dense topic) and that infers *sparse documents* of topic mixtures (i.e., documents each is mixture of a limited number of topics instead of all topics as in a dense document).

We developed a *Fully Sparse Topic Model* (FSTM) [6] that can be viewed as a simplified variant of LDA [14] (without Dirichlet prior) and PLSA [21] without observable documents. These facts allow only few topics to contribute to a document. This relaxation allows us to infer really sparse topic proportions of documents with Frank-Wolfe algorithm [18]. No employment of Dirichlet prior over topics enables us to learn models of low complexity, i.e., sparse models.

A topic model often assumes that a given corpus is composed from K topics, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K)$, and each document is a mixture of those topics. Example models include PLSA, LDA and many of their variants. Under those models, each document has another latent representation. Such latent representations of documents can be inferred once those models have been learned previously.

Definition 1 (Topic proportion) *Consider a topic model $\mathfrak{M}$ with K topics. Each document $\boldsymbol{d}$ will be represented by $\boldsymbol{\theta} = (\theta_1, ..., \theta_K)^t$, where θ_k indicates the proportion that topic k contributes to $\boldsymbol{d}$, and $\theta_k \geq 0, \sum_{k=1}^{K} \theta_k = 1$. $\boldsymbol{\theta}$ is called* topic proportion *(or latent representation) of $\boldsymbol{d}$.*

Definition 2 (Inference) *Consider a topic model $\mathfrak{M}$ with K topics, and a given document $\boldsymbol{d}$. The inference problem is to find the topic proportion that maximizes the likelihood of $\boldsymbol{d}$ under the model $\mathfrak{M}$.*

For some applications, it is necessary to infer which topic contributes to a specific emission of a term in a document. Nevertheless, it may be unnecessary for many other applications.

The *task of learning* in FSTM is to learn all topics $\boldsymbol{\beta}$, given a corpus $\mathcal{C}$. We use EM scheme to iteratively learn the model as usual. Specifically, we repeat the following two steps until convergence:

E-step: do inference for each document of $\mathcal{C}$;

M-step: maximize the likelihood of $\mathcal{C}$ with respect to $\boldsymbol{\beta}$.

The *task of inference* in FSTM is done by connecting with concave optimization based on a proved lemma, and allows us to seamlessly use the Frank-Wolfe algorithm for inference. An appropriate adaptation to the Frank-Wolfe algorithm [18] results in an inference algorithm for FSTM.

Table 1: Data for experiments. $\bar{n}$ is the average number of different terms in a document.

Data	M	Testing size	V	Classes	$\bar{n}$
AP	2,021	225	10,473	0	135
KOS	3,087	343	6,906	0	103
Grolier	23,044	6,718	15,276	0	80
Enron	35,875	3,986	28,102	0	96
20Newsgroups	15,935	3,993	62,061	20	80
Webspam	350,000	350,000	16,609,143	2	3,728

Table 2: Results of learning FSTM from Webspam

Number of topics	1000	2000
Time per EM iteration	28 minutes	65 minutes
EM iterations to reach convergence	17	16
Topic sparsity (compared with dense models)	0.0165 (60 times smaller)	0.0114 (87 times smaller)
Document sparsity (compared with dense models)	0.0054 (185 times smaller)	0.0028 (357 times smaller)
Storage for the new representation ($\boldsymbol{\theta}$) (compared with the original corpus)	31.5 Mb (757 times smaller)	33.2 Mb (718 times smaller)
Average length of topic proportions, $\bar{s}$ (compared with dense representations)	5.4 (185 times smaller)	5.6 (357 times smaller)

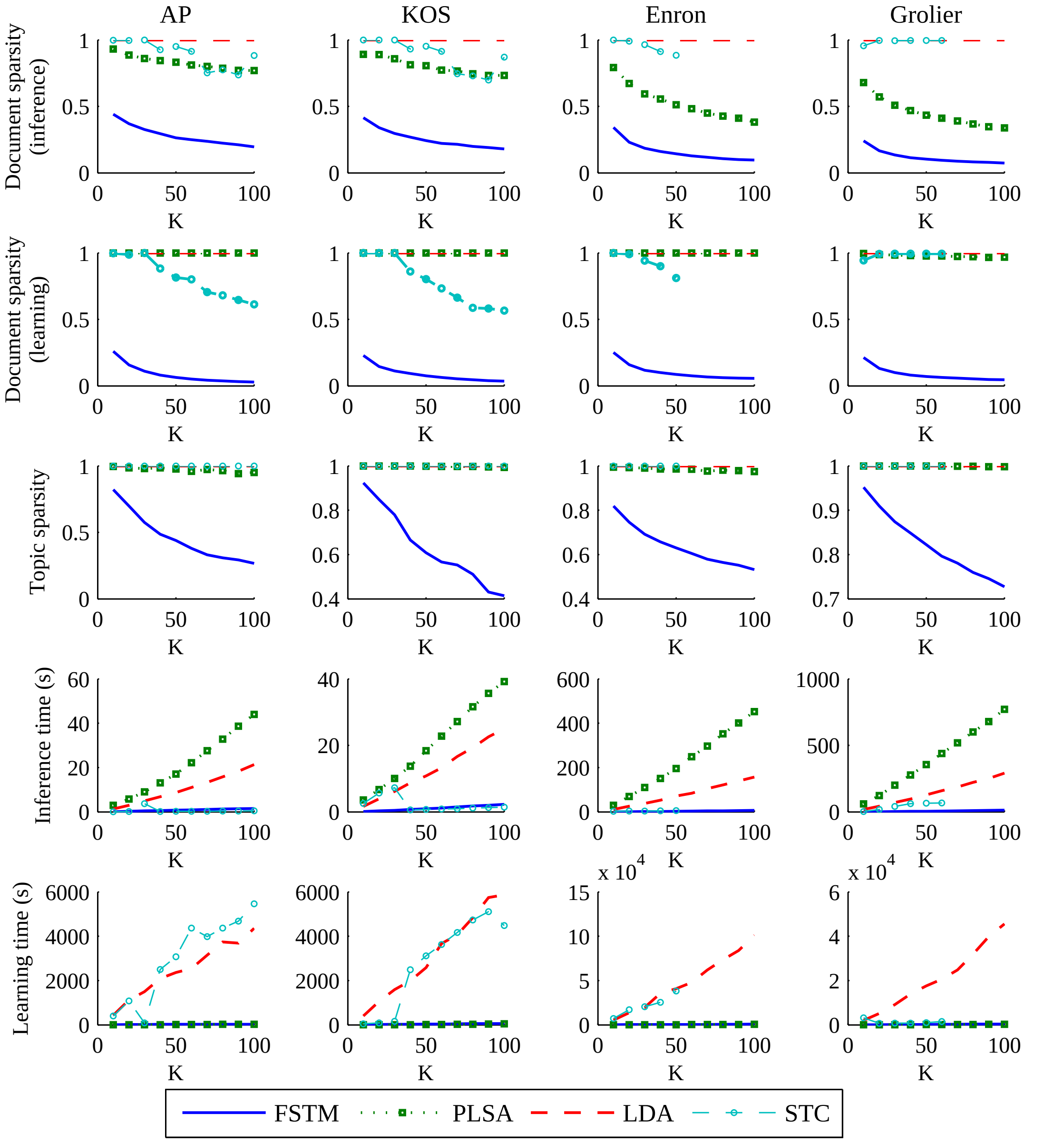

Figure 1: Comparative experimental results of 4 topic models (FSTM, LAD, PLSA and STC [30]) as the number K of topics increases. The lower line the better method. For STC, there was a memory problem when dealing with Enron and Grolier for large K (e.g., when $K = 70$, STC has to solve a optimization problem with more than 20 millions of variables, and hence cannot be handled in a personal PC with 6Gb memory.) Hence we could not do experiments for such large K's.

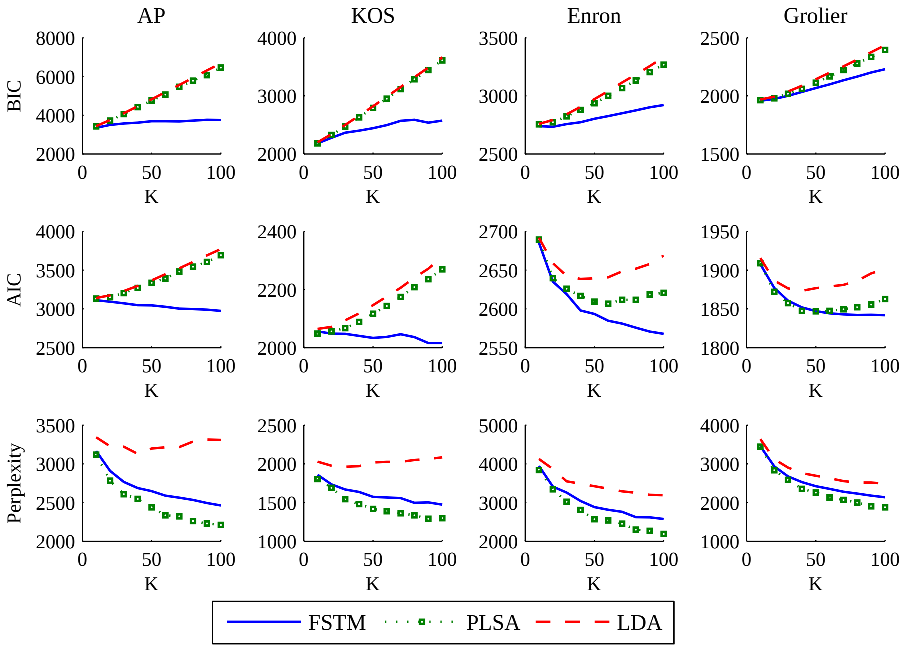

Figure 2: Quality of three models as the number of topics increases. The lower line the better method.

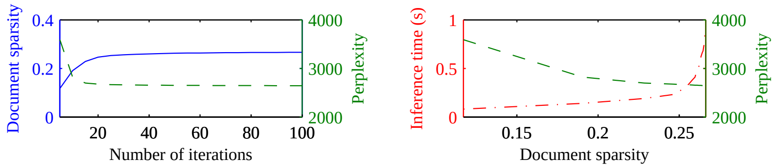

Figure 3: Illustration of trading off sparsity against quality and time. More iterations imply better quality, but probably denser topic proportions. Inference was done on AP, where FSTM had been learned with 50 topics.

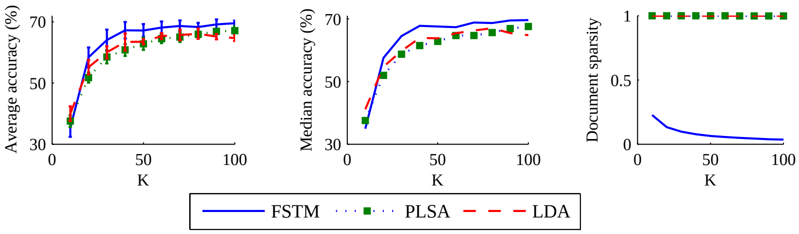

Figure 4: Classification on 20Newsgroups with dimensionality reduction. On the right shows how sparse the learned topic proportions are. We see that FSTM used few features to represent documents while PLSA and LDA used most features.

2.1.2 Discussion on FSTM

In this work, we present our initial step towards resolving the mentioned four large-scale settings. Our attempts attack the two fundamental issues mentioned before by seeking fast inference algorithms and sparse models.

Our first contribution is the introduction of *Fully Sparse Topic Model* (FSTM). Loosely speaking, FSTM is a simplified variant of LDA when relaxing the Dirichlet priors over hidden topics and over hidden topic proportions of documents. It is also a simplified variant of PLSA when removing the observed variable associated with each document. Nevertheless, FSTM has some following attractive properties:

- Inference is done by the Frank-Wolfe algorithm [18] which converges at a linear rate to the optimal solutions. The inference algorithm allows us to swiftly recover sparse topic proportions. Further, it provides a principled way to directly trade off sparsity of solutions against inference quality and running time.
- Learning of topics amounts to multiplication of two sparse matrices. Hence topics are often very sparse. The sparsity level can be directly controlled.
- The complexity of the learning algorithm is near independent of the dimensionality.
- There is an implicit prior over topic proportions, though no explicit employment of priors. Such a prior can help FSTM avoid overfitting.

For the first time in the topic modeling literature, FSTM is the model that couples the two interesting properties: near dimension-free learning algorithm, and ability to trade off sparsity of solutions against inference quality. The near independence of dimensionality implies that FSTM provides an almost optimal answer to the setting (c). It also implies that there exists a near dimension-free algorithm for doing dimensionality reduction (DR), since topic modeling is an approach to DR. These properties are crucial for dealing with data of extremely high dimensions. We hope that our results open a motivation for future studies to seek dimension-free algorithms for other problems.

The ability of FSTM to learn sparse topics and to infer sparse latent representations of documents allows us to save substantially memory for storage. Combined with a linear inference algorithm, FSTM overcomes severe limitations of existing probabilistic models and can deal well with the settings (b), (c), and (d). Fast learning of topics and fast inference of documents also enable us to deal well with the setting (a).

Table 3: Large-scale classification on Webspam. Though reducing the dimensionality drastically, the quality of classification is still comparably maintained.

Data	Dimensions	Storage	Accuracy	Classified by
Original Webspam	16609143	23.3 Gb	*99.15%*	BMD [Yu et al. 2012]
When reducing dimensionality with FSTM				
1000 topics	1000	31.5 Mb	*98.877%*	FSTM + Liblinear
2000 topics	2000	33.2 Mb	*99.146%*	FSTM + Liblinear

To see more advantages of FSTM over existing models, we report some theoretical characteristics of some closely related models in Table 4.

Our second contribution is a distributed architecture for learning FSTM from large data. We employ both distributed scheme for data and task parallelism. Warm-start is further used to speed up learning, while keeping comparable quality. All of these provide a scalable learning algorithm that can handle very large corpora. In particular, we successfully learned a topic model with more than 33 billions of latent variables, from a large corpus with a vocabulary of 16 millions terms. This is the largest model that has been learned in the literature up to now.

A side contribution is the introduction of the Frank-Wolfe algorithm for doing inference in admixture topic models. This algorithm has many attractive properties such as having linear convergence rate, swiftly recovering sparse solutions, providing a way to directly trade off sparsity of solutions against quality and time. Those properties are essential in order to resolve large-scale settings. Moreover, such properties make the Frank-Wolfe algorithm more attractive than traditional inference methods such as folding-in [21], variational methods [20], Gibbs sampling [24], [27].

Extensive experiments show that FSTM works well in practice. It significantly outperforms many models in terms of learning time, inference time, model complexity, and sparsity of latent representations of documents. The predictive power is observed to be comparable with other models. In terms of generalization on unseen data, FSTM often does better. Qualitative performance of FSTM is also observed in application to classification, for both small and very large data.

We theoretically and experimentally show that FSTM can provide provably good solutions. It requires modestly few arithmetic operations, linear in the length of the document to be inferred or in the number of topics. The learning algorithm has very low complexity which does not depend on the size V of the vocabulary. Further, we

Table 4: Theoretical comparison of 8 topic models: FSTM, PLSA, LDA, FTM (Williamson, 2010), SparseTM (Wang, 2009), STC (Zhu, 2011), SRS (Shashanka, 2007), RLSI (Wang, 2011). V is the vocabulary size, K is the number of topics, $\bar{n}$ is the average length of documents. $\bar{K}$ is the average number of topics to which a term has nonzero contributions, $\bar{K} \leq K$. '-' denotes *'no'* or *'unspecified'*; '✓' means *'yes'* or *'taken in consideration'*.

Model	FSTM	PLSA	LDA	FTM	SparseTM	STC	SRS	RLSI
Document sparsity	✓	-	-	✓	-	✓	✓	-
Topic sparsity	✓	-	-	-	✓	-	✓	✓
Sparsity control	direct	-	-	indirect	indirect	indirect	indirect	indirect
Trade-off:								
sparsity vs. quality	✓	-	-	-	-	-	-	-
sparsity vs. time	✓	-	-	-	-	-	-	-
Dimension-free learning	✓	-	-	-	-	-	-	-
Inference complexity	$O(\bar{n}.\bar{K}+K)$	$O(\bar{n}.K)$	$O(\bar{n}.K)$	-	-	$O(\bar{n}.K)$	$O(\bar{n}.K)$	$O(V.\bar{K}^2+K^3)$
Storage for topics	$V.\bar{K}$	$V.K$	$V.K$	-	-	$V.K$	$V.\bar{K}$	$V.\bar{K}$
Auxiliary parameters	0	0	0	0	0	3	2	2

can easily trade off quality of solution against sparsity and inference time. Existing topic models do not own these interesting properties.

Theorem 1. *Consider FSTM with K topics, and a document $\boldsymbol{d}$. Let C_f be defined as $C_f = -\frac{1}{2}\sup_{\boldsymbol{y},\boldsymbol{z}\in\Delta;\tilde{\boldsymbol{y}}\in[\boldsymbol{y},\boldsymbol{z}]}(\boldsymbol{y}-\boldsymbol{z})^t.\nabla^2 f(\tilde{\boldsymbol{y}}).(\boldsymbol{y}-\boldsymbol{z})$ for the function $f(\boldsymbol{x}) = \sum_{j\in I_d} d_j \log x_j$. Then algorithm 1 converges to the optimal solution with a linear rate. In addition, after L iterations, the inference error is at most $4C_f/(L+3)$, and the topic proportion $\boldsymbol{\theta}$ has at most $L+1$ non-zero components.*

Theorem 2. *Each iteration of our algorithm requires only $O(n.\bar{K}+K)$ arithmetic operations, where $\bar{K}$ is the average number of topics to which a term has non-zero contributions, $\bar{K} \leq K$, and $n = |I_d|$. Overall, after L iterations, algorithm 1 requires $L.O(n.\bar{K}+K)$.*

The benchmark datasets shown in Table 1 were used in our experiments. Figure 1 summarizes the comparative experimental results about sparsity and time. Document sparsity is used to see sparsity level of latent representations discovered by those models. Figure 2 and Figure 3 show the quality of three models on four corpora. Figure 4 and Tables 2-3 are about large scale learning.

2.1.3 Fast online inference for topic models [5]

Topic models such as LDA or FSTM face a challenge to analyze very large text collections. To this end, we recently developed the new method with three novel contributions [5]: (1) a proof for the tractability of the MAP estimation of topic mixtures under certain conditions that might fit well with practices, even though the

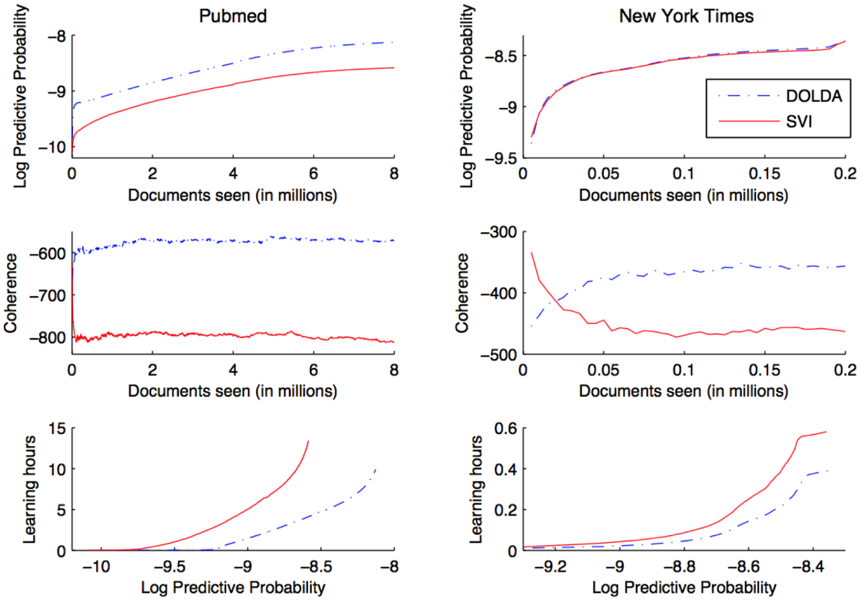

Figure 5: Performance of DOLDA and SVI on two large corpora when learning 100-topic LDA. The higher the better for Predictive Probability and Coherence, whereas lower is better for Learning hours. The last row shows how long the two methods reach to the same generalization level.

problem is known to be intractable in the worse case; (2) the provably fast algorithm OFW (Online Frank-Wolfe) for inferring topic mixtures; (3) the dual online algorithm DOLDA (Dual online LDA) for learning LDA at a large scale. We show that OFW converges to some local optima, but under certain conditions it can converge to global optima. The discussion of OFW is general and hence can be readily employed to accelerate the MAP estimation in a wide class of probabilistic models. From extensive experiments we find that DOLDA can achieve significantly better predictive performance and semantic quality, with lower run-time, than stochastic variational inference. Further, DOLDA enables us to easily analyze text streams or millions of documents. Based on the result and experience with DOLDA, we are carrying out a more challenging step: an dual online for FSTM.

Our experiments aim to see how well DOLDA learns in comparison with SVI (stochastic variational inference proposed by Hoffman in 2013). Figure 5 presents the results on two corpora. One can easily observe that as seeing more documents, both DOLDA and SVI reached to better predictiveness levels with a fast rate. For Pubmed, DOLDA performed signicantly better than SVI even just after seeing a few thousands

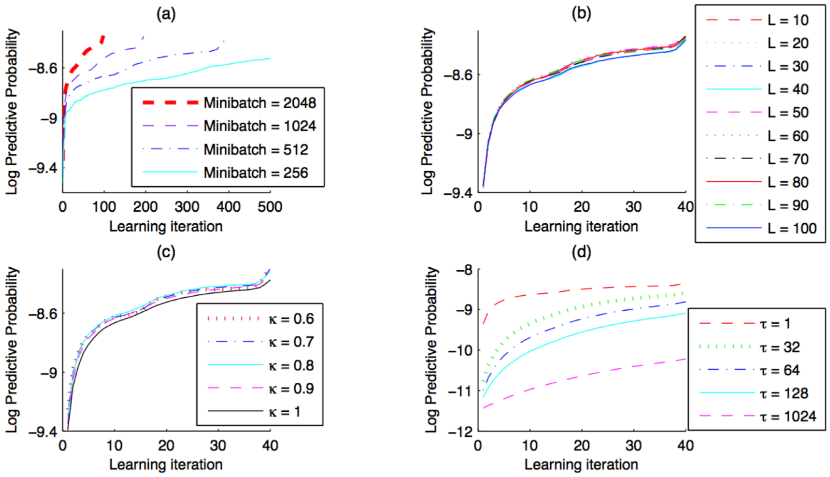

Figure 6: Sensitivity of DOLDA when changing parameters. (a) Change the minibatch size when fixed $\{\kappa = 0.9, \tau = 1, L = 50\}$. (b) Change the number L of iterations for OFW when fixed $\{\kappa = 0.9, \tau = 1\}$. (c) Change the forgetting rate κ when fixed $\{\tau = 1, L = 50\}$. (d) Change τ when fixed $\{\kappa = 0.9, L = 50\}$. The minibatch size in the cases of (b), (c), (d) is 5000. These experiments were done on New York Times, with K = 100 topics.

of documents. DOLDA often reached at the same generalization level (measured by log predictive probability) as SVI within a much less runtime. SVI often needed much more time and data to reach the same prediction level as DOLDA. This demonstrates the goodness of our algorithm.

We investigated the effects of the parameters on the performance of DOLDA. The parameters include: the forgetting rate κ, τ, the number L of interations for OFW, and the minibatch size. Inappropriate choices of those parameters might affect significantly the performance of DOLDA. To see the effect of a parameter, we changed its values in a finite set, but fixed the other parameters. Results of our experiments are depicted in Figure 6.

We also want to see the convergence rate of OFW, inference time, and stability. To this end, we took the 100-topic LDA as a fixed model which has been learned by SVI previously from New York Times; and then we did inference on individual testing documents by OFW and VB. Both methods were allowed 100 iterations to do inference on a document. Results are depicted in Figure 7.

2.2 Supervised dimension reduction (SDR) [1]

The task of *supervised dimension reduction* (SDR) is to find a new space of K dimensions which preserves the predictiveness of the response/label variable Y. Loosely speaking, predictiveness preservation requires that projection of data points onto the new space should preserve separation (discrimination) between classes in the original space, and that proximity between data points is maintained. Once the new space is determined, we can work with projections in that low-dimensional space instead of the high-dimensional one. Our approach is based on topic modeling.

We propose a novel framework which consists of two phases. Loosely speaking, the first phase tries to find an initial topical space, while the second phase tries to utilize label information and local structure of the training data to find the discriminative space. The first phase can be done by employing an unsupervised topic model such as LDA, FSTM, and hence inherits scalability of unsupervised models. Label information and local structure in the form of neighborhood will be used to guide projection of documents onto the initial space, so that inner-class local structure is preserved and inter-class margin is widen. As a consequence, the discrimination property is not only preserved, but likely made better in the final space.

Consider a corpus $\mathcal{D} = \{\boldsymbol{d}_1, ..., \boldsymbol{d}_M\}$ consisting of M documents which are composed from a vocabulary of V terms. Each document $\boldsymbol{d}$ is represented as a vector of term frequencies, i.e. $\boldsymbol{d} = (d_1, ..., d_V) \in \mathbb{R}^V$, where d_j is the number of occurrences of term j in $\boldsymbol{d}$. Let $\{y_1, ..., y_M\}$ be the class labels assigned to those documents, respectively.

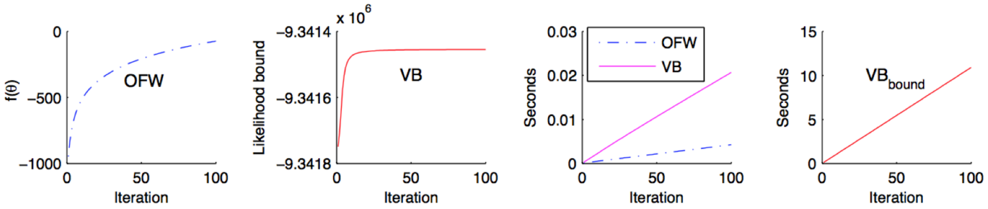

Figure 7: Convergence and inference time of OFW and VB as the number of iterations increase. The first two subplots show how fast OFW and VB maximize their objective functions, while the last two subplots show how long they took. The last subplot shows how long VB did inference when the lower bound of $Pr(d|\beta, \alpha, \eta)$ was used to check convergence. Note that VB did hundreds of times faster than VB_{bound}, i.e., checking bounds for convergence in VB requires intensive time.

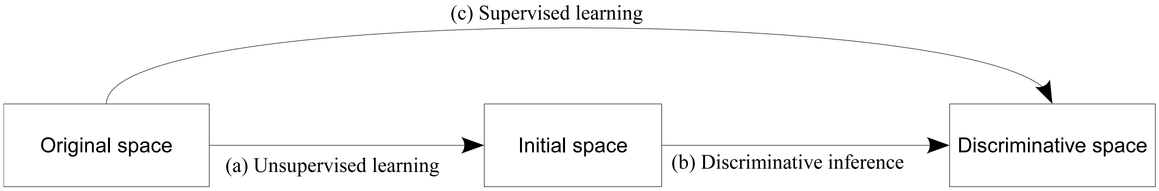

Figure 8: Sketch of approaches for SDR. Existing methods for SDR directly find the discriminative space, which is known as supervised learning (c). Our framework consists of two separate phases: (a) first find an initial space in an unsupervised manner; then (b) utilize label information and local structure of data to derive the final space.

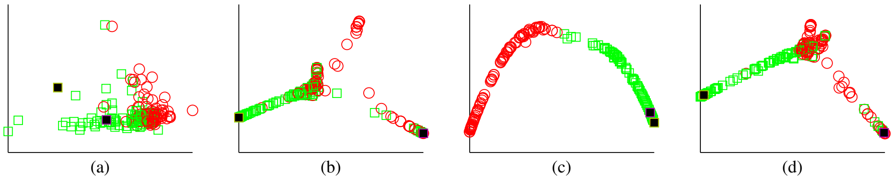

Figure 9: Laplacian embedding in 2D space. (a) data in the original space, (b) unsupervised projection, (c) projection when neighborhood is taken into account, (d) projection when topics are promoted. These projections onto the 60-dimensional space were done by FSTM and experimented on 20Newsgroups. The two black squares are documents in the same class.

Figure 8 depicts graphically this framework, and a comparison with other one-phase methods. Note that we do not have to design entirely a learning algorithm as for existing approaches, but instead do one further inference phase for the training documents. Details of our framework are presented in Algorithm 1.

2.2.1 Why is the framework good?

We theoretically elucidate the main reasons for why our proposed framework is reasonable and can result in a good method for SDR. In our observations, the most important reason comes from the choice of the objective for inference. Inference with that objective plays three crucial roles to preserve or make better the discrimination property of data in the topical space.

- The first role is to preserve inner-class local structure of data. This is a result of using the additional term $\frac{1}{|N_d|}\sum_{\boldsymbol{d}'\in N_d} L(\widehat{\boldsymbol{d}'})$. Since nearest neighbors N_d are

Algorithm 1 Two-phases framework for SDR

```
Phase 1: learn an unsupervised model to get $K$ topics $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K$. Let $\mathfrak{A} = span\{\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K\}$ be the initial space.
Phase 2: (finding discriminative space)
(2.1) for each class $c$, select a set $S_c$ of topics which are potentially discriminative for $c$.
(2.2) for each document $\boldsymbol{d}$, select a set $N_d$ of its nearest neighbors which are in the same class as $\boldsymbol{d}$.
(2.3) infer new representation $\boldsymbol{\theta}_d^*$ for each document $\boldsymbol{d}$ in class $c$ using the Frank-Wolfe algorithm with the objective function $f(\boldsymbol{\theta}) =$
      $$\lambda.L(\widehat{\boldsymbol{d}}) + (1-\lambda).\frac{1}{|N_d|}\sum_{\boldsymbol{d}' \in N_d} L(\widehat{\boldsymbol{d}'}) + R.\sum_{j \in S_c} \sin(\theta_j),$$
      where $L(\widehat{\boldsymbol{d}})$ is the log likelihood of document $\widehat{\boldsymbol{d}} = \boldsymbol{d}/||\boldsymbol{d}||_1$; $\lambda \in [0,1]$ and $R$ are nonnegative constants.
(2.4) compute new topics $\boldsymbol{\beta}_1^*, ..., \boldsymbol{\beta}_K^*$ from all $\boldsymbol{d}$ and $\boldsymbol{\theta}_d^*$. Finally, $\mathfrak{B} = span\{\boldsymbol{\beta}_1^*, ..., \boldsymbol{\beta}_K^*\}$ is the discriminative space.
```

selected within-class only, doing projection for $\boldsymbol{d}$ in step (2.3) is not intervened by documents from outside classes. Hence within-class local structure would be better preserved.

- The second role is to widen the inter-class margin, owing to the term $R\sum_{j \in S_c} \sin(\theta_j)$. The projection of $\boldsymbol{d}$ is encouraged to be close to the topics which are potentially discriminative for class c. Hence projection of class c is preferred to distributing around the discriminative topics of c. Increasing the constant R implies forcing projections to distribute more densely around the discriminative topics, and therefore making classes farther from each other. Figure 9(d) illustrates the benefit of this second role.

- The third role is to reduce overlap between classes, owing to the term $\lambda L(\widehat{\boldsymbol{d}}) + (1-\lambda)\frac{1}{|N_d|}\sum_{\boldsymbol{d}' \in N_d} L(\widehat{\boldsymbol{d}'})$ in the objective function. This is a very crucial role that helps the two-phases framework works effectively. Explanation for this role needs some insights into inference of $\boldsymbol{\theta}$.

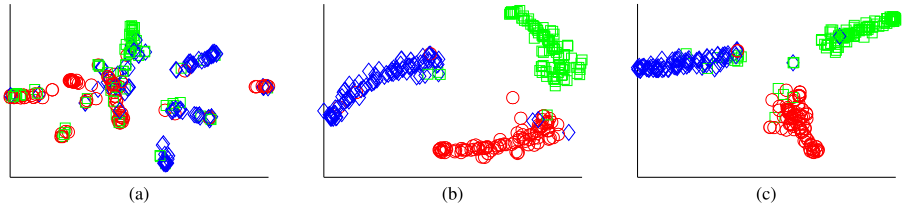

Figure 10: Projection of three classes of 20newsgroups onto the topical space by (a) FSTM, (b) FSTM^c, and (c) MedLDA. FSTM did not provide a good projection in the sense of class separation, since label information was ignored. FSTM^c and MedLDA actually found good discriminative topical spaces, and provided a good separation of classes.

2.2.2 Class separation and classification quality

Separation of classes in low-dimensional spaces is our first concern. A good method for SDR should preserve inter-class separation of data in the original space. Figure 10 depicts an illustration of how good different methods are.

The framework SDR was demonstrated to exploit well label information and local structure of the training data to find a discriminative low-dimensional space. Generality and flexibility of our framework was evidenced by adaptation to three unsupervised topic models, resulted in PLSA^c, LDA^c, and FSTM^c for supervised dimension reduction. These methods can perform qualitatively comparably with the state-of-the-art method, MedLDA. In particular, FSTM^c performed significantly best and can often achieve more than 10% improvement over MedLDA. Meanwhile, FSTM^c consumes substantially less time than MedLDA does. These results show that our framework can inherit scalability of unsupervised models to yield competitive methods for supervised dimension reduction.

The resulting methods (PLSA^c, LDA^c, and FSTM^c) are not limited to discrete data. They can work also on non-negative data, since their learning algorithms actually are very general. Hence in this paper, we contributed methods for not only discrete data but also non-negative real data. The code of these methods is available at www.jaist.ac.jp/∼s1060203/codes/sdr/

There is a number of possible extensions to our framework. First, one can easily modify the framework to deal with multilabel data. Second, the framework can be modified to deal with semi-supervised data. A key to these extensions is an appropriate utilization of labels to search for nearest neighbors, which is necessary for our framework. Other extensions can encode more prior knowledge into the objective function for inference. In our framework, label information and local neighborhood are encoded into the objective function and have been observed to work well. Hence, we believe that other prior knowledge can be used to derive good methods.

Of the most expensive steps in our framework is the search for nearest neighbors. By a modest implementation, it requires $O(k.V.M)$ to search k nearest neighbors for a document. Overall, finding all k nearest neighbors for all documents requires $O(k.V.M^2)$. This computational complexity will be problematic when the number of training documents is large. Hence, a significant extension would be to reduce complexity for this search. It is possible to reduce the complexity to $O(k.V.M.\log M)$. Furthermore, because our framework use local neighborhood to guide projection of documents onto the low-dimensional space, we believe that approximation to local structure can still provide good result. However, this assumption should be studied further. A positive point of using approximation of local neighborhood is that computational complexity of a search for neighbors can be done in linear time $O(k.V.M)$.

2.3 Non-negative matrix factorization

2.3.1 Simplicial non-negative matrix factorization [4]

1. sNMF Problems

Mathematically, we can define the NMF problem as follows:

Definition 5 (NMF): *Given a dataset consisting of M N-dimension vectors $X = [X_1, X_2, ..., X_M] \in R_+^{M \times N}$, where each vector presents a data instance. NMF seeks to decompose X into a product of two nonnegative factorizing matrices F and G, where $F = [F_1, ..., F_M] \in R_+^{M \times K}$ and $G = [G_1, ..., G_K] \in R_+^{K \times N}$ are coefficient matrix and latent component matrix, respectively, $X \approx FG$.*

We proposed a new NMF formulation, called *Simplicial Non-negative Matrix Factorization* with expected properties. The technical details can be found in papers [3]. We assume that each instance is a convex combination of the latent components obtained by adding a new *simplicial constraint* into NMF. Hence, we have:

Definition 6 (Simplicial NMF): *Simplicial NMF is NMF where each instance X_m is a convex combination of the latent components $X_m \approx \sum_{k=1}^{K} F_{mk} G_k$ and $\sum_{k=1}^{K} F_{mk} = 1$ for all m.*

By adding this new constraint, we have associated a probabilistic model with NMF problem, in which each instance is a probabilistic distribution over the latent components and represented as a convex combination of latent components. In other words, this convex combination provides explicitly the extent of contribution of each latent component, while other formulations of NMF do not have. Moreover, regarding to geometry meaning, each instance is projected as a point on the simplex of latent components. This projection is called instance inference. As a result, we obtained significant properties: sparsity, convexity, fast computation, clear interpretability, distributability and parallelizability.

To control the quality of NMF, various cost functions are employed. The cost functions $f(X||FG)$ often contain two parts: The first part is a divergence function that measures the distance between original coordinates (X) and inverted coordinates (FG); and the second one is possibly regularizations and constraints to control sparsity or orthogonality.

Recently, there are numerous divergence functions, including squared *Euclidean* distance, *KL*-divergence, α-divergence, β-divergence, *IS* divergence, and *Bregman* divergence, etc. A chosen divergence mainly depends on the data type and its properties. The two most popular divergences are widely used in numerous applications:

- Squared *Euclidean* distance: $D(x||y) = ||x - y||_2^2 = \sum_i (x_i - y_i)^2$
- *KL*-divergence: $D(x||y) = \sum_i x_i.log\frac{x_i}{y_i} - x_i + y_i$, where x and y are positive vectors.

With these divergence functions, we have two basic problems of simplicial NMF (sNMF):

- sNMF with squared *Euclidean* distance $J(X||FG) = \sum\limits_{m=1}^{M} D(X_m||F_mG)$, where $D(X_m||F_mG) = ||X_m - F_mG||_2^2$
- sNMF with *KL*-divergence $J(X||FG) = \sum\limits_{m=1}^{M} D(X_m||F_mG)$, where $D(X_m||F_mG) = \sum\limits_{n=1}^{N}(X_{mn}.log\frac{X_{mn}}{[F_mG]_n} - X_{mn} + [F_mG]_n)$; $X, F, G \geq 0$; $\sum\limits_{k=1}^{K} F_{mk} = 1$ for all m.

The following algorithms for learning and inference are given in the reference.

Algorithm 1: Inference for data instance x

```
Input: Data instance $x$ and latent components $G = \{g_k\}_{k=1}^K$
Output: New coefficient $f$ minimizing $h = ||x - fG||_2^2$
begin
    Choose component $g_k$ closest to $x$;
    Set $f = \mathbf{0}$; $f_k = 1$; and $r = x - g_k$;
    repeat
        Select $k = argmin_{k \in \{1..K\}}[\frac{\partial h}{\partial f}]_k$;
        $\alpha = r(g_k - x)^T / ||g_k - fG||_2^2$;
        $\alpha = min(\alpha, 1)$;
        $\alpha = max(\alpha, max(-1, -\frac{f_k}{1-f_k}))$;
        if $\alpha == 0$ then
            break;
        Set $r = x - \alpha g_k - (1-\alpha)(x - r)$;
        Set $f = (1-\alpha)f$ and $f_k = f_k + \alpha$;
    until False;
```

Algorithm 2: Inference for data instance x

```
Input: Data instance $x$ and latent components $G = \{g_k\}_{k=1}^K$
Output: New coefficient $f$ minimizing $h(f) = \sum_{n=1}^{N}(x_n log \frac{x_n}{[fG]_n} - x_n + [fG]_n)$
begin
    Choose component $g_k$ closest to $x$.;
    Set $f_i = \mathbf{0}$; $f_k = 1$;
    repeat
        Select $k = argmin_{i \in \{1..K\}}[\frac{\partial h}{\partial f}]_k$;
        $\alpha = argmin_{\alpha \in [0,1]} h(\alpha g_k + (1-\alpha) fG)$;
        Set $f = (1-\alpha)f$ and $f_k = f_k + \alpha$;
    until convergence condition satisfied;
```

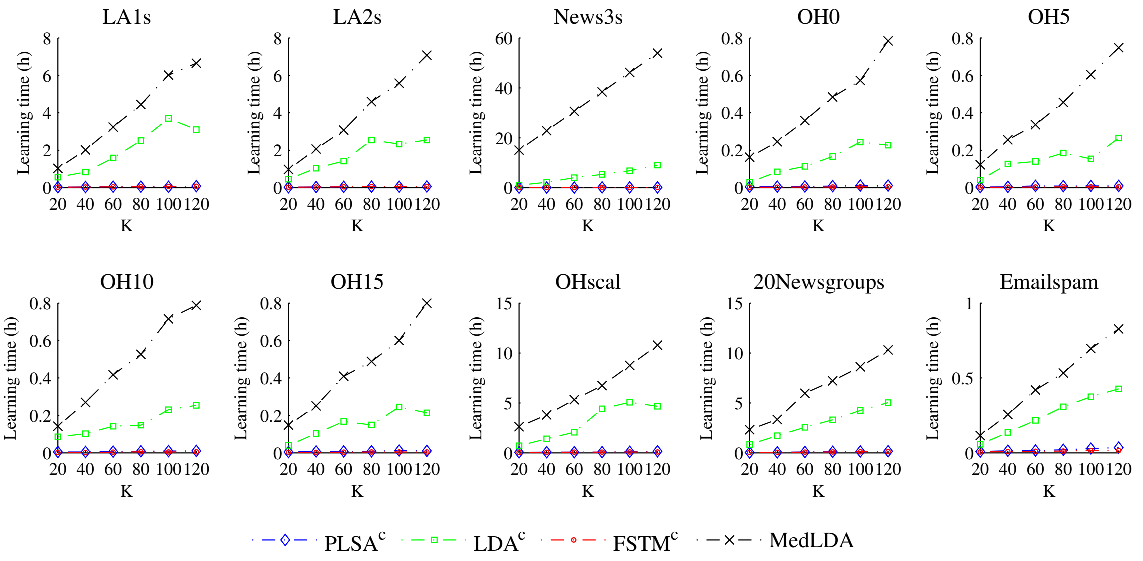

Figure 11: Necessary time to learn a discriminative space, as the number K of topics increases. FSTM^c and PLSA^c often performed substantially faster than MedLDA. As an example, for News3s and $K = 120$, MedLDA needed more than 50 hours to complete learning, whereas FSTM^c needed less than 8 minutes.

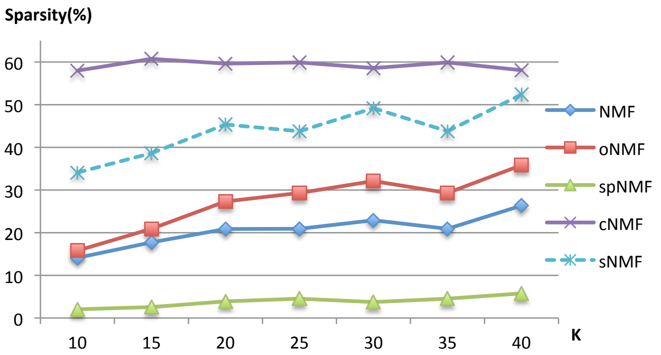

Figure 12: Sparsity of new coefficients for *Euclidean* distance with $K = 30$

2. Complexity

2.1 Complexity for sNMF with Squared Euclidean Distance

Theorem 4. *Consider Algorithm 1 to infer a data instance having N-dimension by K latent components with L iterations. Then its complexity is $O(L[K.S(N)+N])$, where $S(N)$ is a function estimate the number of non-zero elements in latent components and $S(N) \leq N$.*

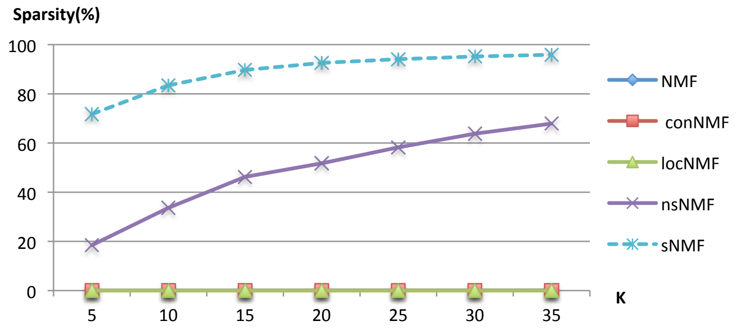

Figure 13: Sparsity of new coefficients for KL-divergence with $K = 30$

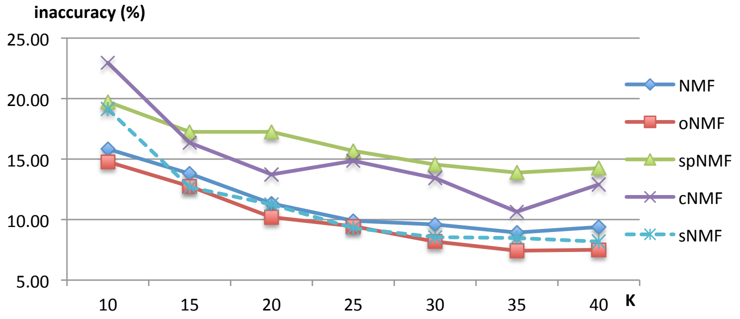

Figure 14: Inaccuracy for Digit Classification

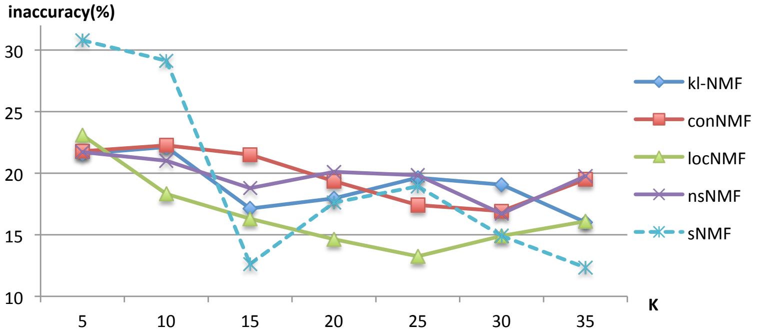

Figure 15: Inaccuracy for Spam Classification

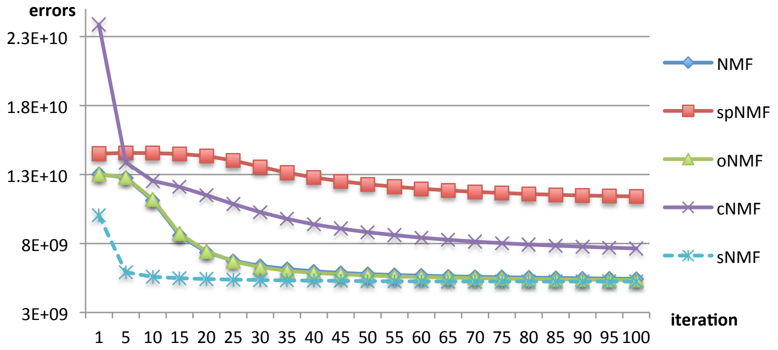

Figure 16: Information Loss for Squared *Euclidean* Distance with $K = 30$

Proof In the Algorithm 1, for each iteration, we have:

$$\frac{\partial h}{\partial f} = 2(fG - X)G^T$$

Hence: $[\frac{\partial h}{\partial f}]_k = 2(fG - X)g_k^T = 2rg_k^T$

Therefore, the complexity of finding out the best coefficient k: $O(K.S(N))$. In addition, the complexity of estimating a is $O(N)$. Overall, the complexity for L iterations is $O(L[K.S(N) + N])$ ■

2.2 Complexity for sNMF with KL-divergence

Theorem 5. *Consider Algorithm 2 to infer a data instance having N-dimension by K latent components with L iterations. Then, its complexity is $O(L[K.S(N)+N.\log\frac{1}{\epsilon}])$.*

In addition, for the learning step with KL-divergence, we employ an approximate algorithm with low complexity:

Theorem 6. *Let f be a twice differentiable convex function over simplex $\triangle$ and denote $C_f = sup_{y,z\in\triangle;\tilde{y}\in[y,z]}(y-z).\bigtriangledown^2 f(\tilde{y}).(y-z)^T$. After l iterations, the Frank-Wolfe algorithm will find an approximate solution x_l with at most $(l+1)$ non-zeros coefficients which satisfy*

$$max_{x\in\triangle} f(x_l) - f(x) \leq \frac{C_f}{l+1}$$

From this theorem, we have the following remarks:

- Convergence rate of inference is linear and the goodness of solutions is bounded, which are crucial in applications.

- Inference depends mostly on complexity of f and $\bigtriangledown f$.

- We can tradeoff easily between sparsity and quality of solutions by stop finding new latent components to optimize the cost function. This property is valid for real applications, which the number of non-zero coefficients is limited.

Theorem 7. *Let consider to learn new latent components after inferring coefficients of data instances. Then, its complexity is $O(M[S(N) + S(K)])$.*

3. Sparse Representation

In order to compare the sparsity of solutions, we compute the percentage of zero coefficients

$$\frac{number\ of\ zero\ coefficients}{number\ of\ coefficients} \times 100$$

The results are highly competitive with other methods. For *Euclidean* distance, although our algorithm's sparsity is only less than cNMF [19] (Figure 12), it has lower information loss and higher performance in classification. In addition, especially for *KL*-divergence, our approach retains the best sparse solutions (Figure 13), while it still has the best result for the other measures.

The results are highly competitive with other methods. For *Euclidean* distance, although our algorithm's sparsity is only less than cNMF [19] (Figure 12), it has lower information loss and higher performance in classification. In addition, especially for *KL*-divergence, our approach retains the best sparse solutions (Figure 13), while it still has the best result for the other measures.

Classification quality is one of measures that evaluates our method's effectiveness as NMF is often considered as a dimension redution technique used widely in classification. In this experiment, we use Random Forest, a robust algorithm for classification. Observing Figures 14 and 15, our method is one of methods with the lowest errors in testing. For *Euclidean* distance and the digit dataset, the result of our method is very close to the best method oNMF [16]. Meanwhile, for KL-divergence and spam dataset, our approach obtains the lowest misclassification with $K = 15$ and $K = 35$.

For dimension reduction, information loss criterion is one of the most important measure. Figures 15 and 16 show that our approach has the lowest information loss.

As a result, we obtained significant properties:

- *Sparsity*: Instance inference is casted as a convex problem over the simplex of latent components by adding the simplicial condition. Furthermore, we can easily control the solution sparsity via greedy approximation algorithms such as Frank-Wolfe algorithm [18].

- *Convexity*: Obviously, inferring an instance is to find an approximation of the convex combination that is a convex optimization problem [15].

- *Computation*: The instance representation can be considered as a projection on the simplex of the latent components. Hence, the inference based on this projection can be much faster than other formulations because of the simplicial constraint added [15]. In comparison to other formulations, this one has significant computing advantages in the inference of instances, while the learning step is the same with the previous basic formulations because they solve the same optimization problem.

- *Interpretability*: The new formulation gives a more comprehensible interpretation of the important role of coefficients. Particularly, each data instance is a convex combination of the latent components, in which the sum of coefficients always equals to 1 through NMF. Hence, the important role of the latent components on instances can be concisely represented via values of coefficients. Otherwise, for other formulations, evaluating the contribution of components is forceful because of the lack of constraints between coefficients. Alternatively, a post-processing can be employed to find out the role of the latent components. However, it is independent and inconsistent with learning NMF model.

- *Distributability and parallelizability*: NMF problem contains two sub-problems: inference and learning. The learning problem is the same with other formulations and can be solved by distributed algorithms [17]. Meanwhile, the inference one of our formulation can be solved by a much faster algorithm comparing to the others', and it can be parallelized [15]. This favor is hard to be reached in other formulations.

The cost function is specially determined on the used divergence function. In this paper, we focus on solving this problem with the two most popular divergence functions with squared *Euclidean* distance and *KL*-divergence.

2.3.2 Anti-lopsided algorithm for non-negative least squares [7]

Non-negative least squares problem (NNLS) is one of the most important fundamental problems in numeric analysis and has been widely used in scientific computation and

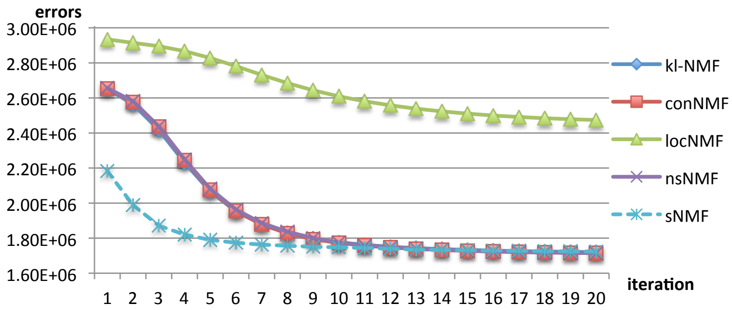

Figure 17: Information Loss for $KL-$divergence with $K = 30$

data modeling. In big data analytics, the current limitations on speed and accuracy of NNLS algorithms remain as typical challenges, and it influences the development of NMF. Targeting to improve NNLS solution, we proposed a fast and robust anti-lopsided algorithm with high accuracy that is totally based on the first order methods. The main idea of our algorithm is to transform the original NNLS problem into an equivalent non-negative quadratic programming problem, which significantly reduces the scaling problem of variables. The proposed algorithm can reach higher accuracy and speed with an exponent convergence rate at least $O(1 - \frac{1}{2||Q||_2})^k$ where $\sqrt{n} \leq ||Q||_2 \leq n$ and n is the dimension size of solutions. The experiments on large matrices clearly show the high performance of the proposed algorithm in comparing to the state-of-the-are algorithms.

We investigate the convergence speed of the square of derivatives $||\bar{f}||_2^2$ in Figure 18 and the difference between the values of objective function and the optimal values $log_{10}(|f_(x_k) - f^*| + 1)$ during the running time, see Figure 19. The results clearly show that our algorithm and algorithm Remarkably, our algorithm comes to the optimal values much more faster than other methods. This favor proves that the anti-lopsided transformation may make iterative methods using the first derivative more effective because it significantly reduces scaling problems of variables.

2.4 Other work on sparse modeling and dimensionality reduction

This part presents some of our work relating to sparse modeling and dimensionality reduction. Some work is not planned and registered to the project, but much inspired from the ideas and methods developed in the project and some directly employed the project's methods.

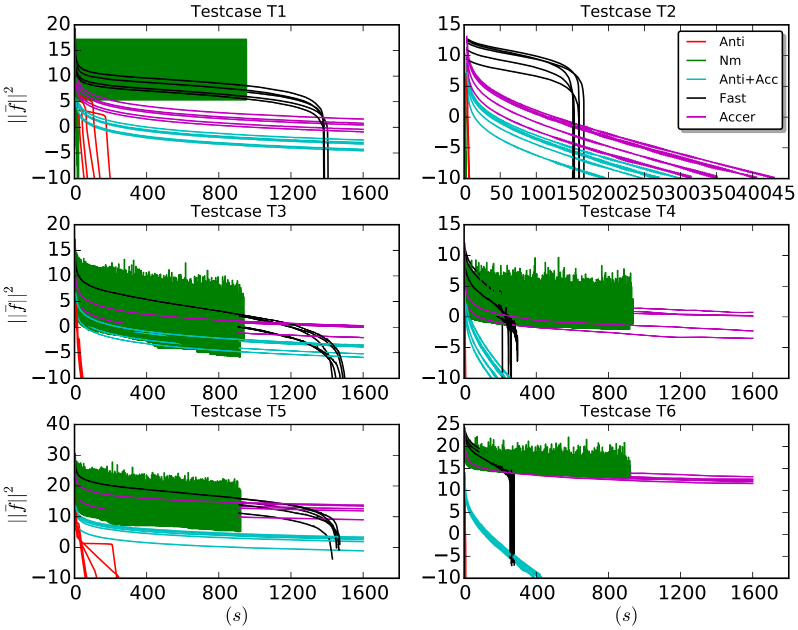

Figure 18: $log_{10}(||\bar{f}||_2^2)$ during runing time

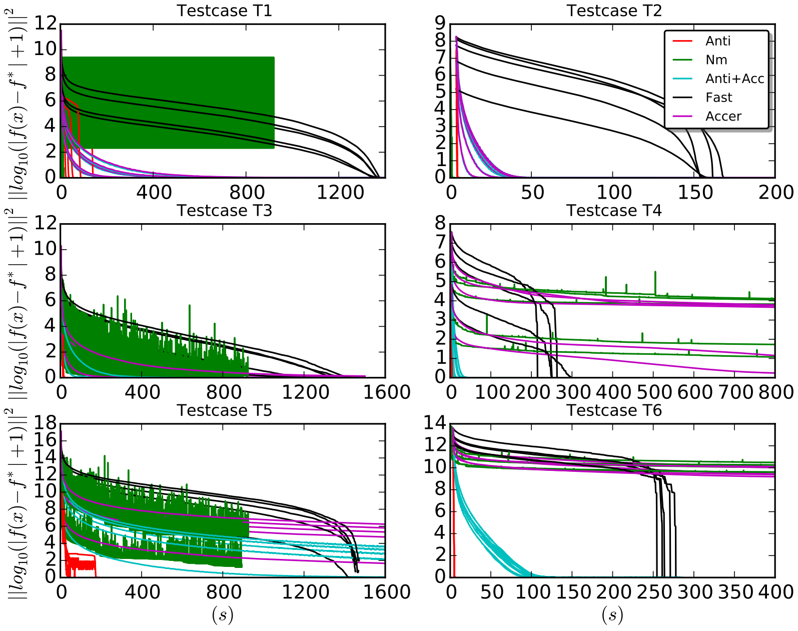

Figure 19: $log_{10}(|f(x_k) - f^*| + 1)$ during running time

2.4.1 Modeling the log-normality data [2]

We investigate two important properties of real data: diversity and log-normality. Log-normality accounts for the fact that data follow the lognormal distribution, whereas diversity measures variations of the attributes in the data. To our knowledge, these two inherent properties have not been paid much attention from the machine learning community, especially from the topic modeling community. In this article, we fill in this gap in the framework of topic modeling. We first investigate whether or not these two properties can be captured by the most well-known Latent Dirichlet Allocation model (LDA), and find that LDA behaves inconsistently with respect to diversity. Particularly, it favors data of low diversity, but works badly on data of high diversity. Then, we argue that these two inherent properties can be captured well by endowing the topic- word distributions in LDA with the lognormal distribution. This treatment leads to a new model, named Dirichlet-lognormal topic model (DLN) [2].

Using the lognormal distribution complicates the learning and inference of DLN, compared with those of LDA. Hence, we used variational method, in which model learning and inference are reduced to solving convex optimization problems. Extensive experiments strongly suggest that (1) the predictive power of DLN is consistent with respect to diversity, and that (2) DLN works consistently better than LDA for datasets whose diversity is large, and for datasets which contain many log-normally distributed attributes. Justifications for these results require insights into the used statistical distributions and will be discussed in the article.

2.4.2 Dimensionality reduction in study of new materials design [3]

In [3]* we worked on application of machine learning in new materials design. For a material with a given hypothesized structural model, the electronic structure, as well as many other physical properties can be predicted by solving the Schrdinger equation. Conventionally, the ground states potential energy of a material is calculated using atomic positions in the hypothesized structure model. By optimizing the ground states potential energy, the optimal structure can be derived. The features of an optimal structure model of materials, as well as its derived physical properties, results in a series of optimizing processes, and in addition has strong multivariate correlations. The task of materials design is to make these correlations clear and to determine a strategy to modify the materials to obtain desired properties. However, such correlations are usually hidden and difficult to uncover or predict by experiments or experience. As a consequence, the design process is currently performed through

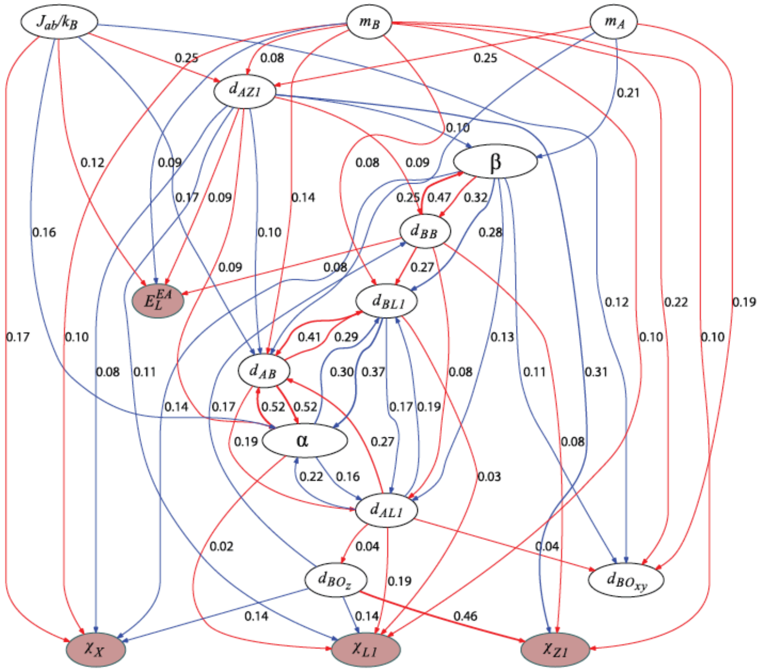

Figure 20: The graph represents all relations between the features. Brown nodes and white nodes indicate independent and dependent features, respectively. Red edges and blue edges indicate positive and negative correlation, respectively. The arrows are from response variables to explanatory variables. The edges are plot with pen-widths in proportion to the values of the corresponding relations.

time-consuming and repetitive experimentation and characterization loops, and to shorten the design process is clearly a big target in materials science.

In an effort to improve on existing techniques, we propose a first principle calculation-based data mining method and demonstrate its potential for a set of computationally designed single molecular magnets with distorted cubane $Mn^{4+}Mn_3^{3+}$ core (Mn4 SMMs). The essential idea of the method is a process consisting of sparse regression model and structure learning with reduction of the relations between features. Figure 18 shows all relations between the features learned from the data, and Figure 19 shows relations between reduced sets of features by using the method.

2.4.3 Dimensionality reduction in biomedicine [8]

In [8] we employed the methods of dimensionality reduction developed by project [1], [6] to analyze the shorten sequences and applied it to find mechanisms of resistance of HCV (hapatitus C virus) typical drugs (interferon and vibavirin). Under the framework of our method, the typical topic models PLSA [20], LDA [14] and our fully sparse topic model FSTM [7] all reached predictive accuracy much higher than SVM with different kernel functions (Table 5).

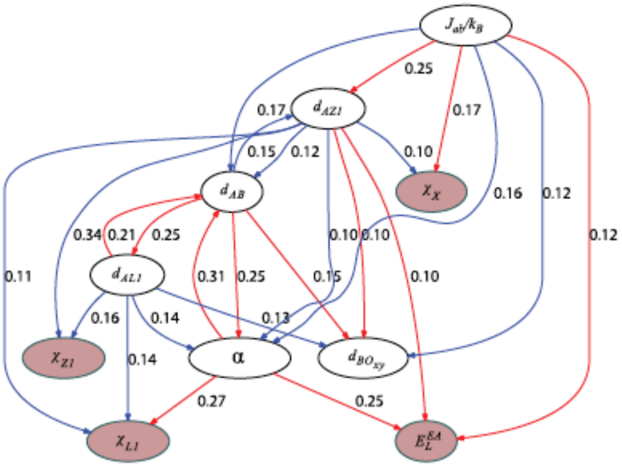

Figure 21: The simplified graph represents the relations between selected features. Brown nodes and white nodes indicate independent and dependent features, respectively. Red edges and blue edges indicate positive and negative correlation, respectively. The arrows are from response variables to explanatory variables. The edges are plotted with pen-widths in proportion to the values of the corresponding relations.

2.4.4 Potential collaboration with research institutions

In our two projects with Fujitsu Hokoriku [11] and with Vietnam National University of Ho Chi Minh City [12], also dimensionality reduction is essential. For predicting the status of IT systems, we have used different methods of reducing the sets of more than 170 log features into about 40 latent features and use them to detect the key factors providing expected properties of the materials.

We are doing joint research on the content of the project with two institutions: (i) Centre for Pattern Recognition and Data Analytics (PRaDA), Faculty of Science, Engineering and Built Environment, Deakin University, Australia; (ii) John von Neumann Institute at Vietnam National University at Ho Chi Minh City, Vietnam.

2.4.5 Software Download

The source codes of FSTM is freely available at http://www.jaist.ac.jp/s1060203/codes/fstm. The source codes of SDR is available at http://www.jaist.ac.jp/s1060203/codes/sdr/

3. List of Publications and Significant Collaborations that resulted from our AOARD supported project

3.1 List of peer-reviewed journal publications

[1] Than, K., Ho, T.B., Nguyen, D.K.: An effective framework for supervised dimension reduction, *Neurocomputing*, Vol. 130, 179-199, 2014.

[2] Than, K., Ho, T.B.: Modeling the diversity and log-normal of data, *Intelligent Data Analysis*, 18(6), 1067-1088, 2014.

[3] Dam, H.C., Pham, T.L., Ho, T.B., Nguyen, T.A., Nguyen, V.C.: Data mining for materials design: A computational study of suingle molecule magnet, *The Journal of Chemical Physics*, 140(4), 2014.

3.2 List of peer-reviewed conference publications

[4] Nguyen, D.K., Than, K., Ho, T.B.: Simplicial Nonnegative Matrix Factorization, *IEEE International Conference on Research, Innovation and Vision for Future*, RIVF 2013, Hanoi, November 10-13, 2013.

[5] Than, K., Doan, T.: Dual online inference for latent Dirichlet allocation, *Asian Conference on Machine Learning*, Supp. Journal Machine Learning Research, Nov. 26-28, 2014.

3.3 Papers currently submitted for review

[6] Than, K., Ho, T.B.: Fully Sparse Topic Models. Submitted to *Journal Machine Learning Research*.

[7] Nguyen, D.K., Ho, T.B.: Anti-lopsided Algorithm for Large-scale Non-negative Least Squares. Submitted to *Journal Ooptimization Letters*.

[8] Le, T.N., Than, K., Kanda, T., Ho, T.B.: Classification and Discriminative Analysis of Short Sequences Using Topic Modeling (completed manuscript to be submitted).

3.4 Conference presentations without papers

[9] Ho, T.B.: Some results on data transformation in machine learning and data mining, *keynote talk at IEEE International Conference on Research, Innovation and Vision for Future*, RIVF 2014, Can Tho, February 25-27, 2015.

[10] Ho, T.B.: On Prediction of siRNA Knockdown Efficiency and Exploitation of Electronic Medical Records, *invited talk, 2nd International Conference on Computational Science and Engineering*, August 21-23, Ho Chi Minh City, 2014.

3.5 List of interactions with industry or with Air Force Research Laboratory scientists or significant collaborations

[11] Project on“Development of data mining techniques to predict anormal status of IT systems from log data” between our lab and Fujitsu Hokuriku, leading by Ho T.B. (JAIST) and Ishii H. (Fujitsu). In the project, we largely employed the results on dimensionality reduction from the AOARD project.

[12] Project on “Establish of basic methodology and generic tools for exploiting electronic medical records” between our lab and Vietnam National University of Ho Chi Minh City, leading by Ho T.B. (JAIST). In the project, we largely employed the results on topic models from the AOARD project.

3.6 Selected list of related work

[13] Bach, F.: Bolasso: Model Consistent Lasso Estimation through the Bootstrap, *ICML 2008*, 2008.

[14] Blei, D., Ng. A.Y., Jordan, M.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, 3, 993-1022, 2003.

[15] Boyd, S.: Alternating direction method of multipliers, NIPS Workshop on Optimization and Machine Learning, 2011.

[16] Choi, S.: Algorithms for orthogonal nonnegative matrix factorization, *IEEE International Joint Conference on Neural Networks*,18281832, 2008.

[17] Chu, C., S.K. Kim, Y.A. Lin, Y. Yu, G. Bradski, A.Y. Ng, K. Olukotun: Map-reduce for machine learning on multicore, *Advances in neural information processing systems*, Vol. 19, 281-, 2007.

[18] Clarkson, K.L.: Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm, *ACM Trans. Algorithms*, 6(63):163, 2010.

[19] Ding, C.H.Q., Tao Li, and M.I. Jordan: Convex and semi-nonnegative matrix factorizations, *Pattern Analysis and Machine Intelligence*, IEEE Transactions on, 32(1):45 55, 2010.

[20] Hoffman, T.: Probabilistic latent semantic indexing, *UAI*, 1999.

[21] Hofmann, T.: Unsupervised Learning by Probabilistic Latent Semantic Analysis, *Machine Learning*, 42, 177-196, 2001.

[22] Koller, D., Friedman, N.: Probabilistic Graphical Models, The MIT Press, 2009.

[23] Meinshausen, N.: Relaxed Lasso, *Computational Statistics and Data Analysis*, 374–393, 2007.

[24] Newman, D., Asuncion, A., Smyth, P., Welling, M.: Distributed algorithms for topic models, *Journal of Machine Learning Research*, 10, 18011828, 2009.

[25] Wang, Q., Xu, J., Li, H., Craswell, N.: Regularized latent semantic indexing: A new approach to large-scale topic modeling, *ACM Trans. Inf. Syst.*, 31 (1):5:15:44, 2013.

[26] Wang, Y.X., Zhang, Y.Z.: Nonnegative matrix factorization: A comprehensive review, *IEEE Trans. on Knowledge and Data Engineering*, 25(6), 13361353, 2013.

[27] Williamson, S., C. Wang, K. A. Heller, D.M. Blei: The ibp compound dirichlet process and its application to focused topic modeling, *International Conference on Machine Learning (ICML)*, 2010.

[28] Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables, *Royal Statistical Society*, 69(1), 49-67, 2007.

[29] Zou, H., Hastie, T., Tibshirani, R.: Sparse Principal Component Analysis, *Journal of Computational and Graphical Statistics*, 15(2), 265-286, 2006.

[30] Zhu, J., Xing, E.: Sparse topical coding, In Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI), 2011.

4. Conclusion

The project has reached its objectives with developed methods for sparse modeling and dimensionality reduction. These methods have been theoretically and experimentally evaluated. This project also opens for us other challenging problems to pursue in coming time.

5. Acknowledgments

We wish to express our sincere thanks to Asian Office of Aerospace Research and Development (AOARD) of Air Force Office of Scientific Research for the support. The funding greatly help us to pursue the significant research on crucial problems of sparse modeling and dimensionality reduction.